**Introduction**

Despite measures taken by the Seattle city traffic department, the number of accidents is still a concern.

In view of the identified accident occurrences, the Seattle City urban planning department, hired our consulting group with the goal of developing an accident prediction system,

The main objectives of the model are**:**

* 1. classify and predict the possibility of a collision causing injuries or physical damage.
  2. determine the conditions and routes most inclined to generate such an accident.

**The Data Available**

The dataset consists of 51 variables and 194.673 observations of collision occurrences from 2004 to 2020

From the 38 variables:

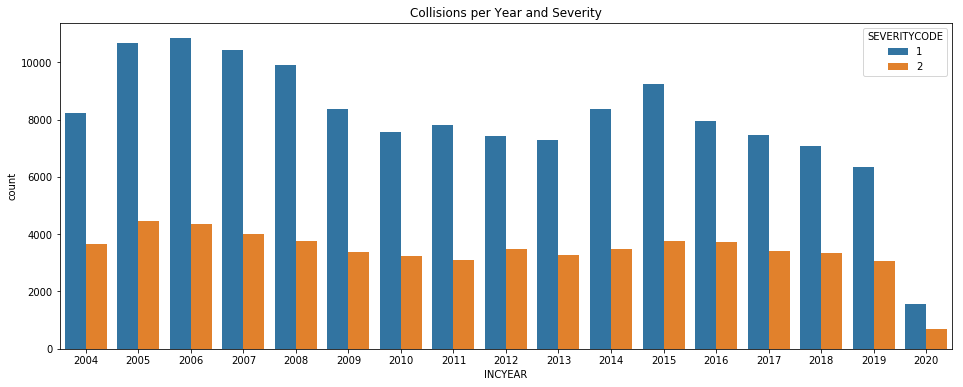
* + - 20 are Objects (Categorical or Boolean)
    - 12 are Integer
    - 4 are Float
    - 2 are Datetime

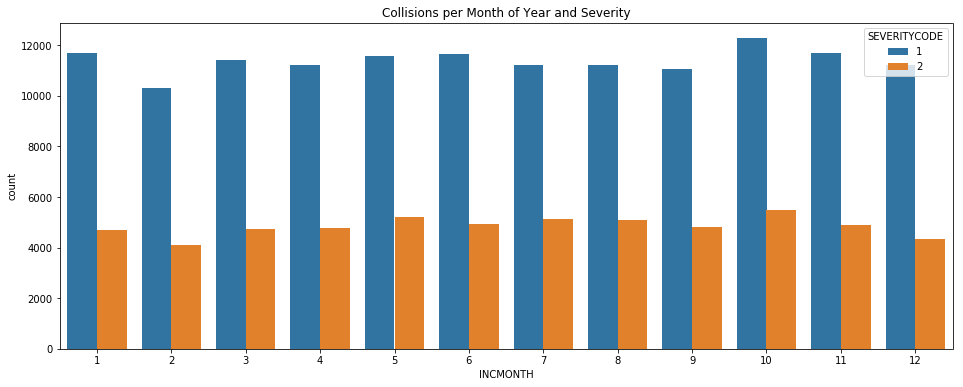
Missing data were found in almost variables and handled as demonstrated bellow

As we can see on the charts on the bellow, the number of collisions has been decreasing consistently. For the analyzed period, the average number of collisions was 11451.

* + Type 1 collisions represent occurrences in which there was only material damage and account for 71% of cases.
  + Type 2 collisions involve physical damage and make up 29% of cases.

Finally, we can see that they are equally distributed among the months of the year. In some cases, this could be important, indicating seasonality. But in our case, we will disregard this information.





**Methodology**

We will use four classification models, in order to determine which model has the best predictive capacity.

**LOGISTIC REGRESSION**: Used to model the probability of a certain class or events occurs, using a cost function.

**DECISION TREE**: A tree-like model of decision, which try to classify the classes, minimize the entropy of the available information.

**KNN**: A non-parametric method, to identify patterns among the data points and classify each occurrence, according to its similarity to others.

**SVM**: A model, which tries to achieve a good separation of classes, trough the construction of a hyperplane that has the largest distance to the nearest training-data point of any class.